



Bio-Rad Laboratories, Inc.
Informatics Division

Application Note

Multi-Technique Spectral Searching in KnowItAll®

Gregory M. Banik, Ph.D., Ty Abshear, and Karl Nedwed

Bio-Rad Laboratories, Inc.
Informatics Division
3316 Spring Garden Street
Philadelphia, PA 19104 USA

Abstract

The use of spectral search software and reference databases in NMR, MS, and IR is a longstanding technique in unknown identification, compound verification, and structure elucidation. Typically, a score or hit quality index (HQI) is calculated to describe the correlation between a spectrum of the unknown compound and spectra of known compounds in reference databases. A new system for multi-technique spectral searching is described that utilizes multi-dimensional analysis of several hit lists resulting from spectral similarity searches performed simultaneously in reference databases for multiple complementary analytical techniques. The multi-dimensional approach permits the optimization of chemical similarity based on several analytical techniques to maximize the chemical knowledge obtained on the unknown compound.

Background

A variety of complementary spectroscopic and chromatographic techniques (Figure 1) are applied in an attempt to answer one of two fundamental issues: compound verification and unknown identification. The former is encountered in quality control environments; the latter in such fields as compound synthesis or natural product identification.

Technique	Information From This Technique
<i>IR & Raman</i>	Which functional groups are present?
<i>NMR</i>	How are carbon and hydrogen atoms bonded?
<i>MS</i>	What is the mass of the molecule and which groups of atoms are present?
<i>UV-Visible</i>	What is the electronic system of the molecule like?
<i>Chromatography</i>	How many different chemicals make up a sample and what is their relative abundance?

Figure 1: Information Obtained from Analytical Techniques

Compound verification and unknown identification progressed significantly with the advent of computer-searchable electronic databases of reference spectra. Algorithmic comparison of the unknown spectra to those in reference databases could provide strong evidence for the compound's identity, or at least clues that could lead to an ultimate identification or confirmation.

In dealing with the search results of reference spectra and their associated chemical structures in a software searching environment, three concepts have become traditional to the point of being dogmatic:

1. *Hit Lists* – A list of similar spectra called a hit list is generated by the software program and displayed in a tabular fashion.
2. *Hit Quality Index* – A hit quality index (HQI) value is associated with each reference spectrum and is a numerical measure of the closeness of fit between the unknown spectrum and each reference spectrum. The higher the HQI value, the closer the unknown spectrum is to the reference spectrum.
3. *Rank Ordering* – The hits are sorted so that those with the highest HQI value are at the top of the list and those with the lowest HQI are at the bottom of the list.

These three concepts are illustrated in Figure 2, which shows a rank-ordered hit list from an IR spectral search.

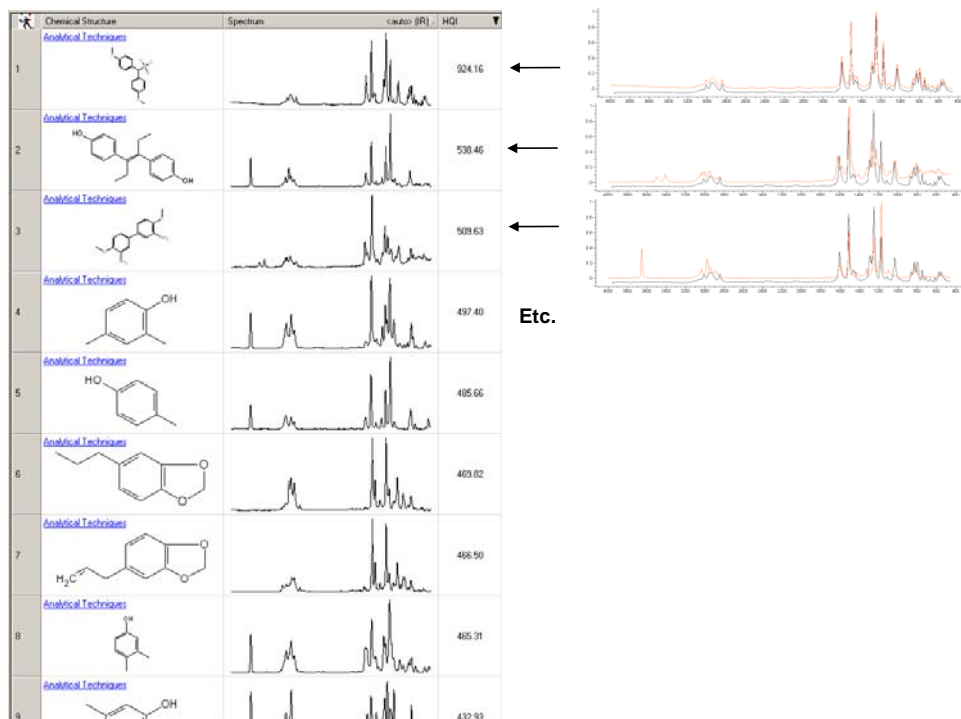


Figure 2: Spectral Hit List and Hit Quality Index (HQI)

Multi-Technique Searching – Theory and Practice

The Informatics Division of Bio-Rad Laboratories, Inc. has introduced a novel and powerful new system for searching multiple spectral techniques simultaneously. Based on the KnowItAll Informatics System, the process is very straightforward: multiple spectra from complementary analytical techniques are used as unknowns to search multiple reference databases containing spectra from the complementary analytical techniques. Two powerful innovations, however, transform the traditional hit lists into a dramatic new approach for visualizing spectral search results.

The first innovation is plotting a hit list from a single spectral search as values on a number line. Illustrated in Figure 3, this simple innovation provides value (albeit limited value) in allowing the relative values of all hits in a hit list to be displayed simultaneously. The second and much more significant innovation is plotting multiple hit lists from multiple spectral searches performed in complementary analytical techniques to be plotted as points in a graph. In this innovation, each point in the graph represents a single compound where the hit quality indices for each reference spectrum in each technique are the coordinates in an N-dimensional spectral space (Figure 4). Points (*i.e.*, compounds from the reference spectral databases) that are closer to the origin will have a lower spectral similarity to the unknown than those farthest from the origin.

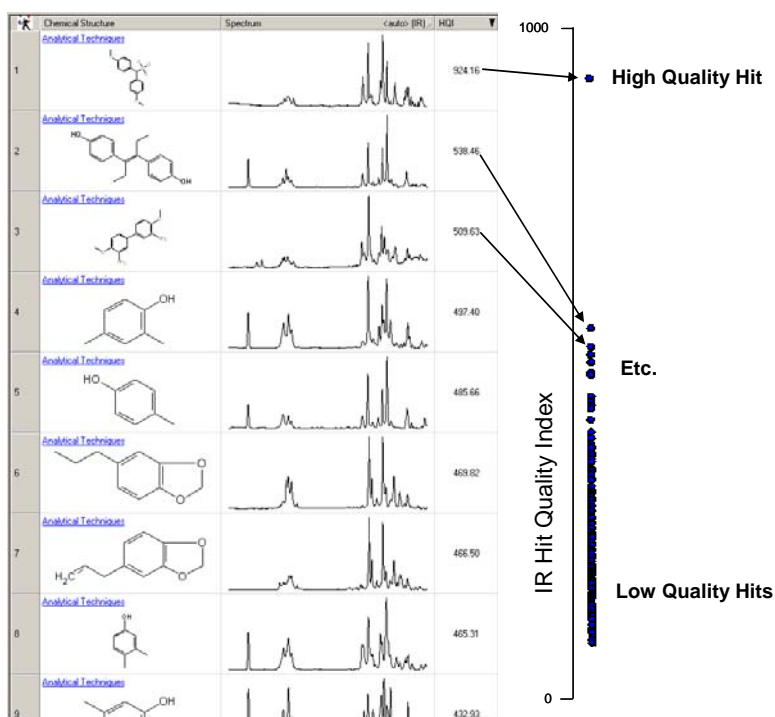


Figure 3 – 1D Plot of HQI Values from One Spectral Technique

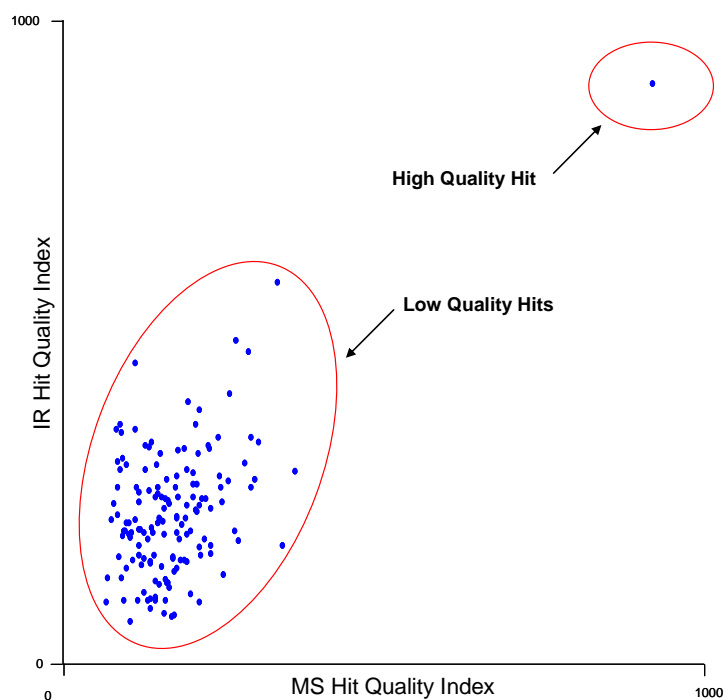


Figure 4 – 2D Plot of HQI Values from Two Spectral Techniques

Materials and Methods

- **Software: KnowItAll® Informatics System (Version 5.0)**
 - IR Searching – Euclidean Distance Algorithm
 - MS Searching – Modified Finnigan Algorithm
 - ¹³C NMR – Peak Search Algorithm
- **IR Databases: HavelItAll® IR (225,000 IR Spectra)**
 - Sadtler (Bio-Rad)
 - Chemical Concepts
- **¹³C NMR Databases: HavelItAll® NMR (360,000 ¹³C NMR)**
 - Sadtler (Bio-Rad)
 - Wolfgang Robien
 - Chemical Concepts
- **MS Databases: HavelItAll® MS (200,000 Mass Spectra)**
 - Chemical Concepts
 - AAFS
 - NIST/EPA/NIH (147,000 spectra) available but not used
- **Test Spectra Sets**
 - IR and MS – NIST WebBook (<http://webbook.nist.gov>)
 - ¹³C NMR – Bio-Rad sample file

Combinations of spectra were searched in the KnowItAll Informatics System (Figure 5). The results from all reference database searches were automatically performed in all spectral techniques simultaneously. The different spectral databases were automatically cross-linked by an exact chemical structure match, and the resulting hit lists were displayed (Figure 6). Spectra were required to exist for all hits, that is, an IR/MS combined search must return a pair of IR and mass spectra linked by exact structure. This automatic procedure eliminates hits where an IR spectrum exists without the corresponding mass spectrum or *vice versa*.

The NIST/EPA/NIH mass spectra reference database was not used as the mass spectra used as test spectra were from the NIST WebBook, and to include them would have biased the study.

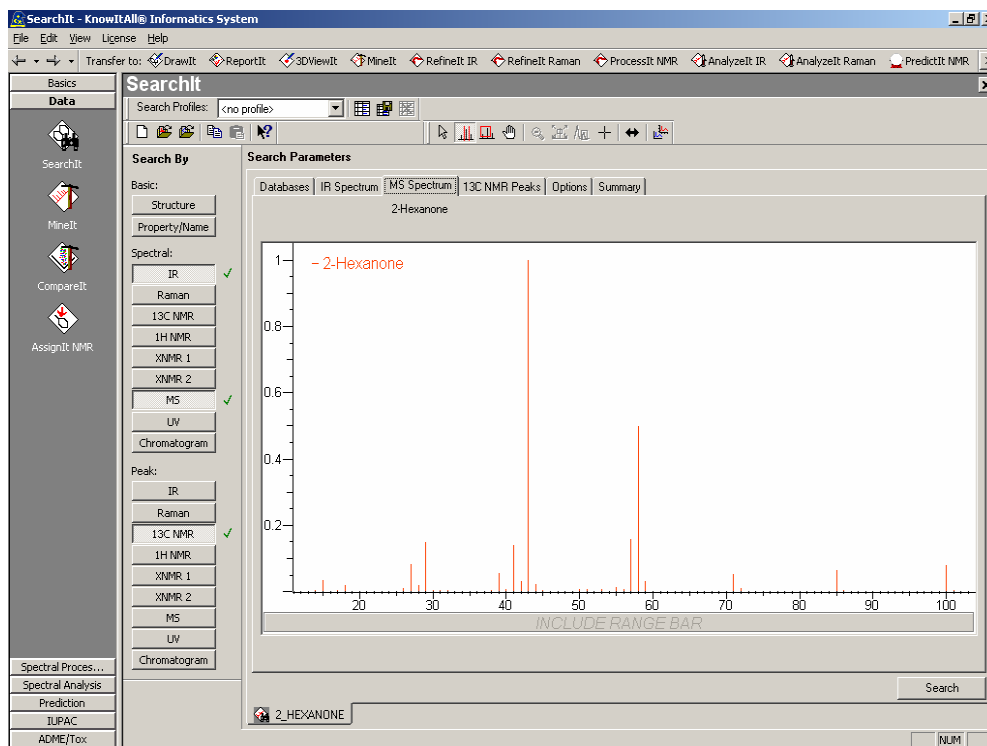


Figure 5 – Defining Multiple Simultaneous Spectral Searches

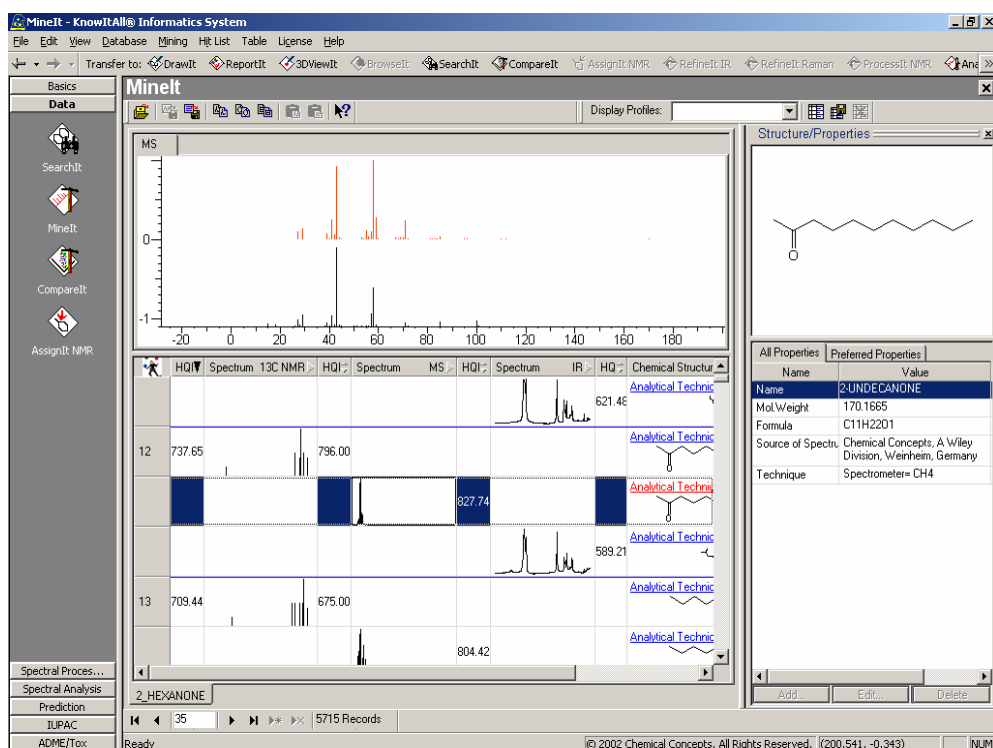


Figure 6 – Multi-Technique Spectral Hitlist (NMR/MS/IR)

The system also automatically removes duplicate spectral hits for the same chemical structure, choosing the database entry with the highest HQI value for each spectral search type. The entire multi-technique hit list was then transferred with a single mouse-click to the data plotting application (CompareIt, Figure 7), where an interactive display allows the user to see the corresponding structures associated with data points by clicking on a single data point or lassoing a group of data points (see arrow, Figure 7).

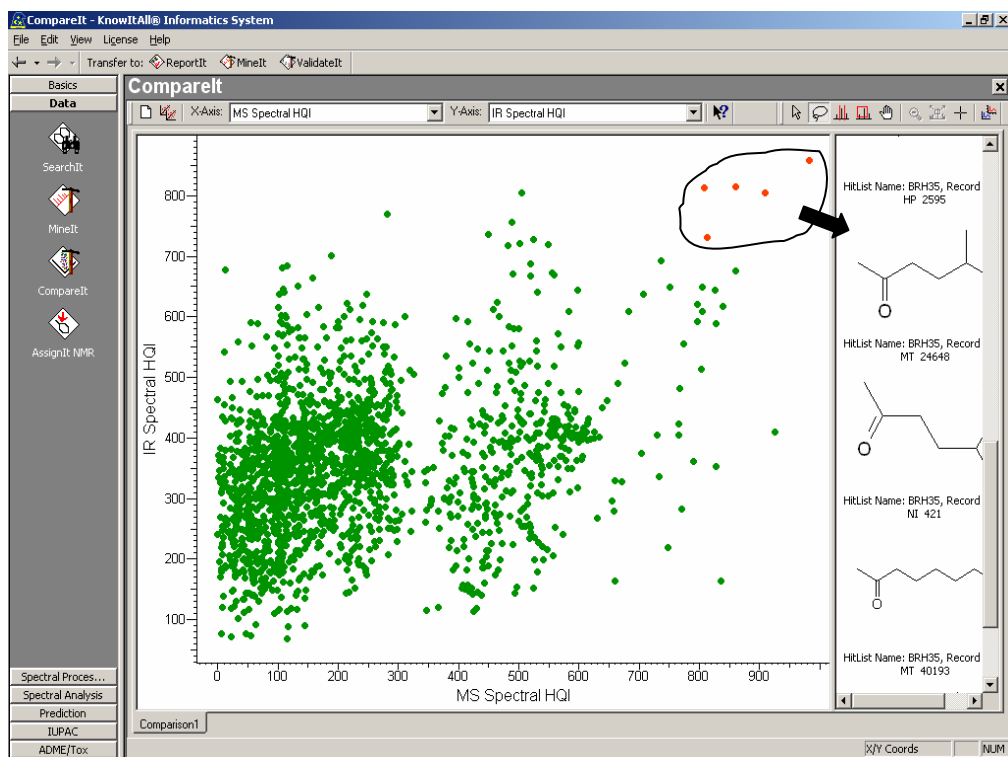


Figure 7 – 2D Spectral HQI Plot (IR/MS)

Results and Discussion

Example 1 – Cholesteryl Acetate

In this example, the IR and mass spectra of cholesteryl acetate (Figure 8) were searched simultaneously. In this case, exact matches for the test spectra were found in the reference spectra databases for both techniques. The resulting 2D plot of MS vs. IR HQI values (Figure 9) shows that neither IR nor MS alone could clearly identify the compound. Using the two techniques in parallel, however, the system clearly identifies this as cholesteryl acetate. Interestingly, a number of related esters of cholesterol (and cholesterol itself) cluster in the upper right of the graph. There is clearly a high degree of correlation between the *spectral space* and the *chemical space*.

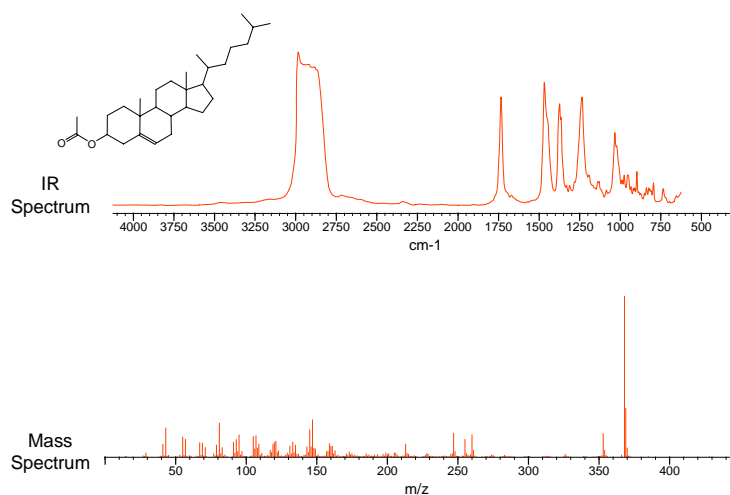


Figure 8 – Example 1: Cholesteryl Acetate

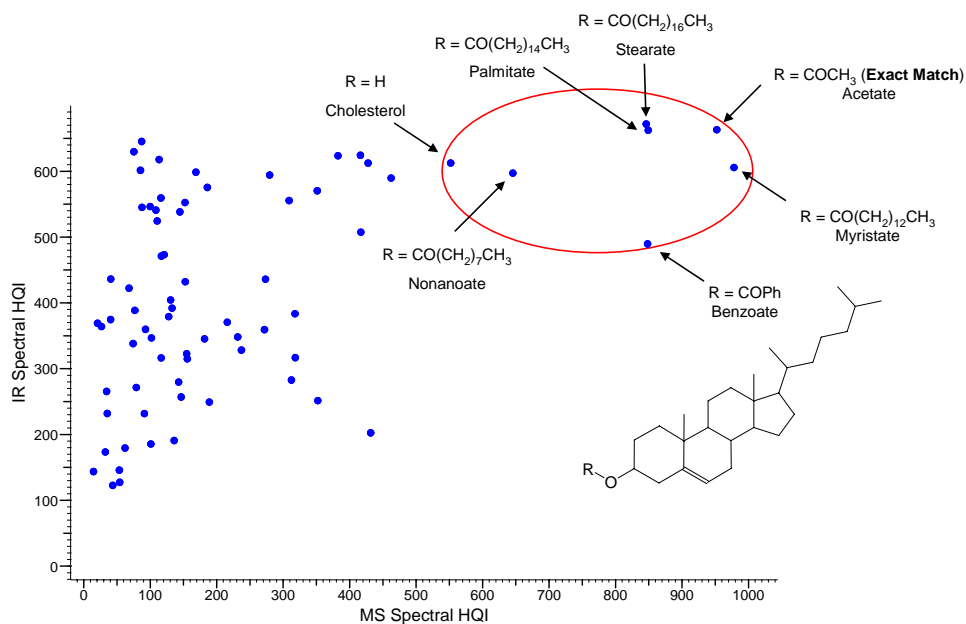


Figure 9 – Example 1 Results: 72 Spectra Pairs (MS/IR)

Example 2 – 1,1,4,4-Tetraphenyl-1,3-Butadiene

In this example, the IR and mass spectra of 1,1,4,4-tetraphenyl-1,3-butadiene (Figure 10) were searched simultaneously. In this case, exact matches for the test spectra were found in the reference spectra databases for both techniques. The resulting 2D plot of MS vs. IR HQI values (Figure 11) shows that in this case, either IR or MS alone could have identified the compound. Using the two techniques in parallel, however, the system degree of separation between the exact match and the rest of the hit list becomes glaringly obvious.

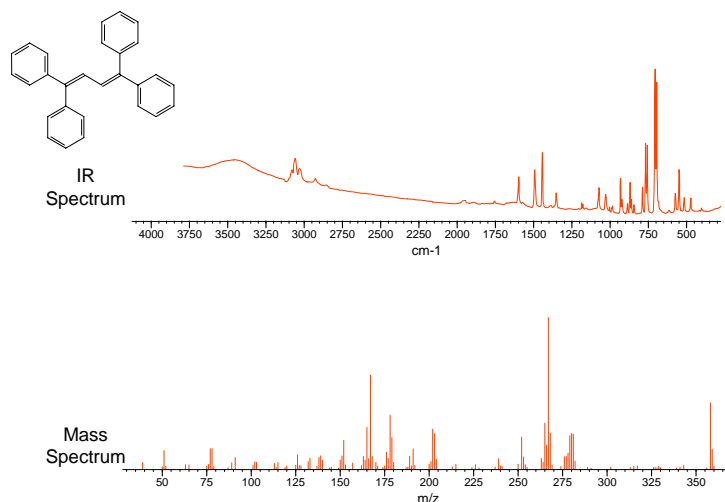


Figure 10 – Example 2: 1,1,4,4-Tetraphenyl-1,3-butadiene

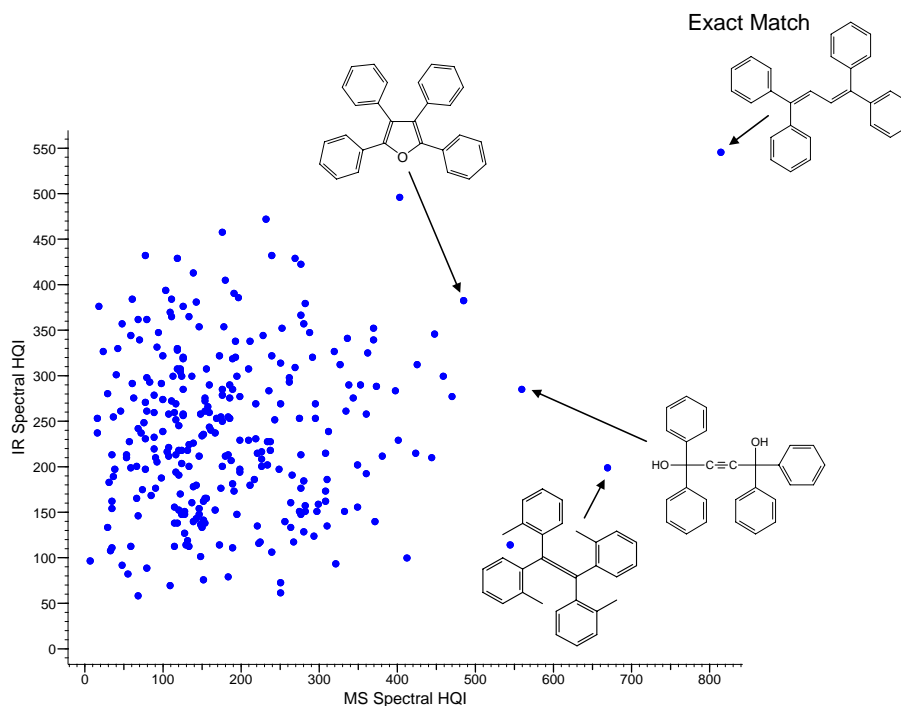


Figure 11 – Example 2 Results: 287 Spectra Pairs (MS/IR)

Example 3 – Phthalic Acid, Butyl Cyclohexyl Ester

In this example, the IR and mass spectra of n-butyl cyclohexyl diester of phthalic acid, (Figure 12) were searched simultaneously. In this case, no exact matches for the test spectra were found in the reference spectra databases. The resulting 2D plot of MS vs. IR HQI values (Figure 13), however, shows an interesting cluster of compounds in the upper right-hand corner of the plot. All of these 11 compounds are phthalic acid diesters. Again, there is clearly a high degree of correlation between the *spectral space* and the *chemical space* in this example.

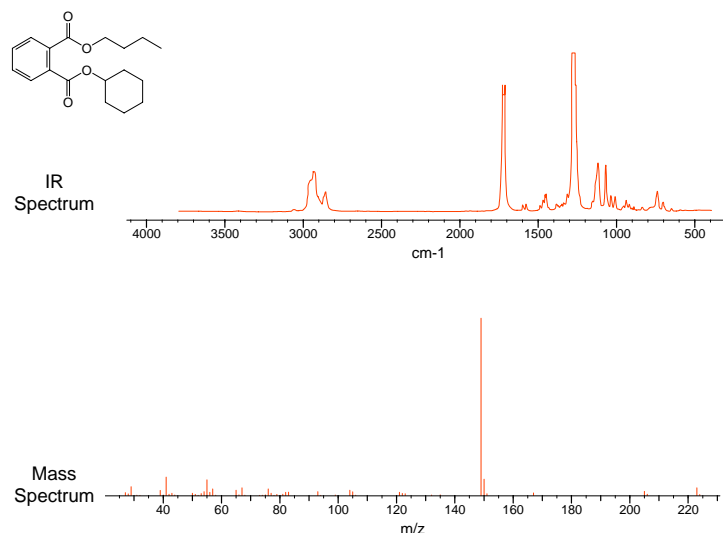


Figure 12 – Example 3: Phthalic Acid, Butyl Cyclohexyl Ester

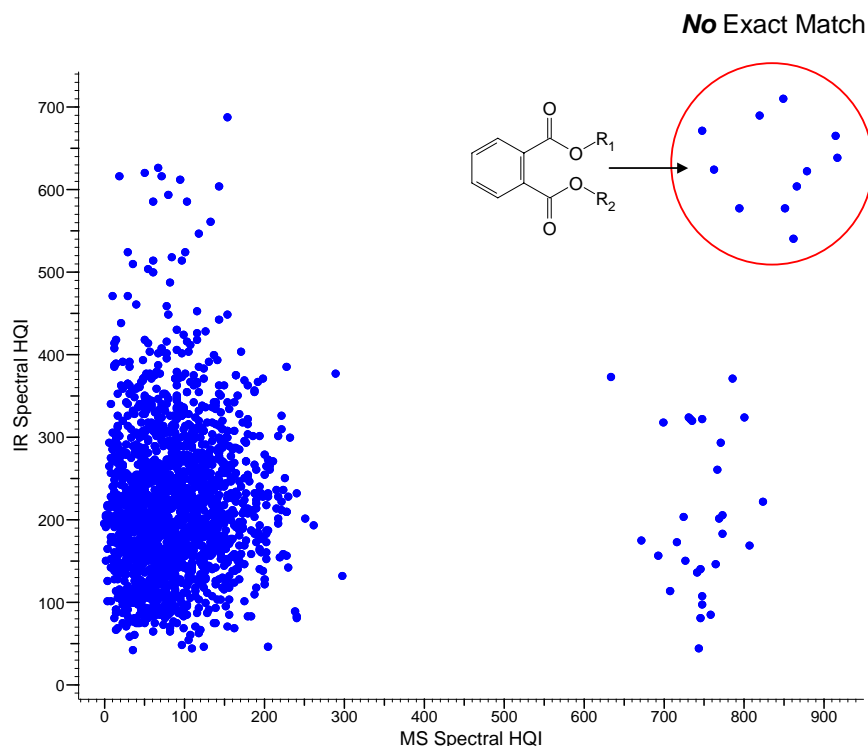


Figure 13 – Example 3 Results: 1,821 Spectra Pairs (MS/IR)

Example 4 – 2-Hexanone

In this example, the ^{13}C NMR, IR and mass spectra of 2-hexanone (Figure 14) were searched simultaneously. In this case, exact matches for the test spectra were found in the reference spectra databases for all three techniques. The resulting 2D plots of ^{13}C NMR vs. IR HQI values (Figure 15), MS vs. ^{13}C NMR HQI values (Figure 16), and MS vs. IR HQI values (Figure 17) show that neither ^{13}C NMR, IR, nor MS alone could clearly identify the compound (*Note: the discrete banding in Figures 15 and 16 is an artifact of the NMR peak search algorithm, which matches a discrete number of peaks in the unknown to entries in the database*). Using the three techniques in parallel, however, the system clearly identifies this as 2-hexanone. Again, a number of related compounds cluster in the upper right of the graph, likewise indicating a high degree of correlation between the *spectral space* and the *chemical space*.

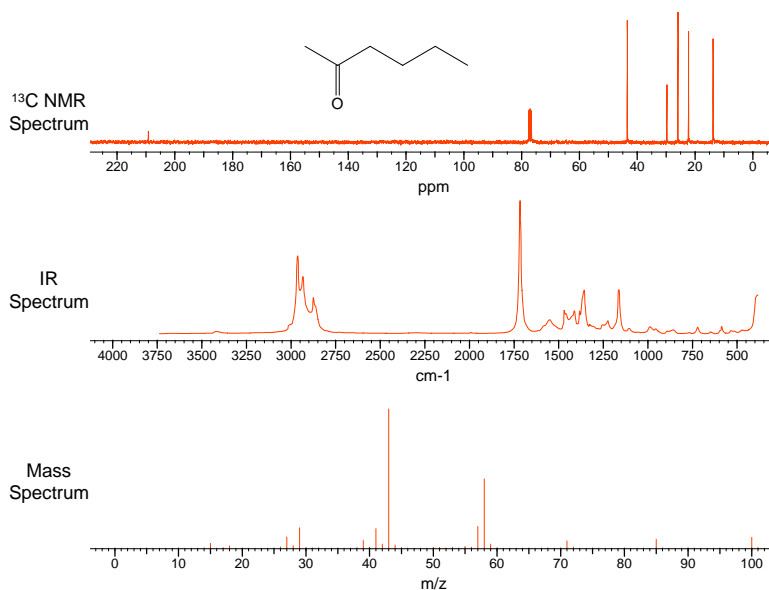


Figure 14 – Example 4: 2-Hexanone

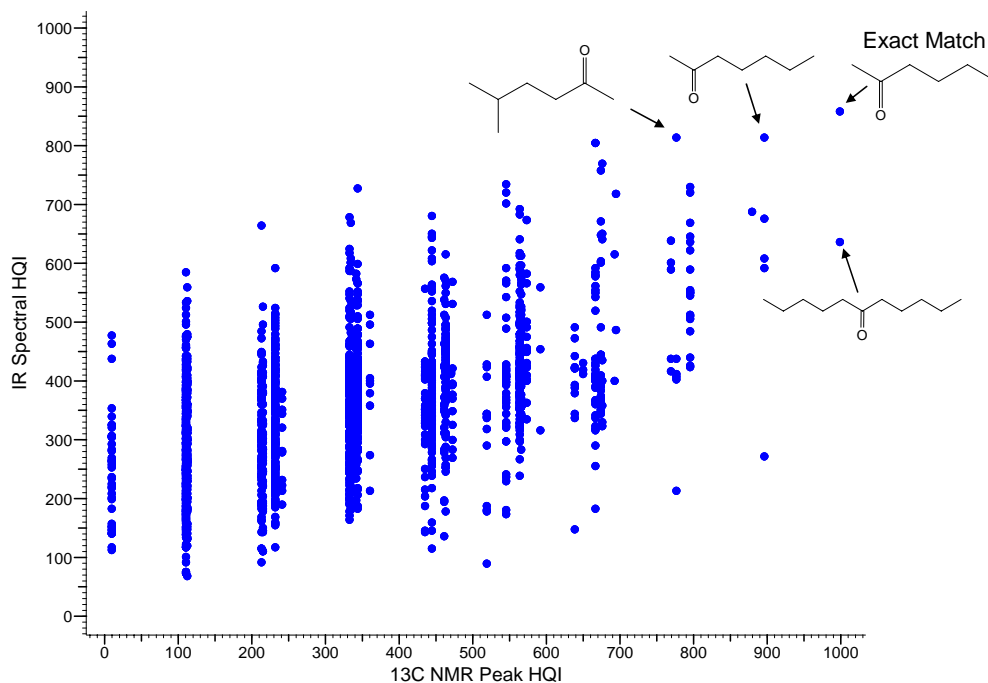


Figure 15 – Example 4 Results: 5,715 Spectra Pairs (NMR/IR)

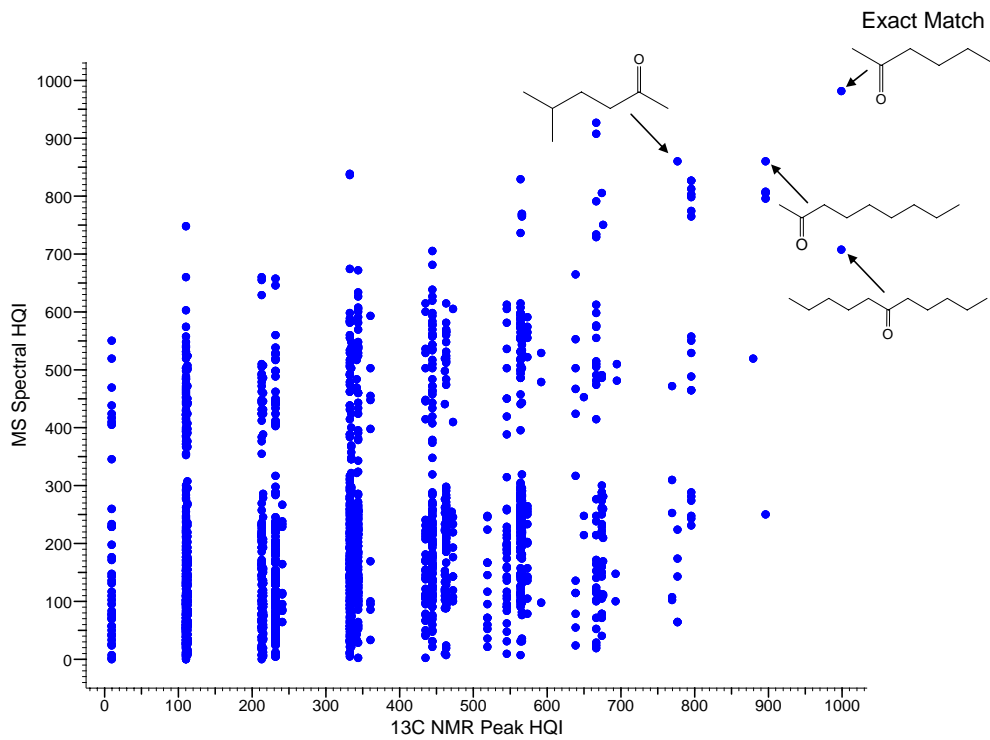


Figure 16 – Example 4 Results: 5,715 Spectra Pairs (NMR/MS)

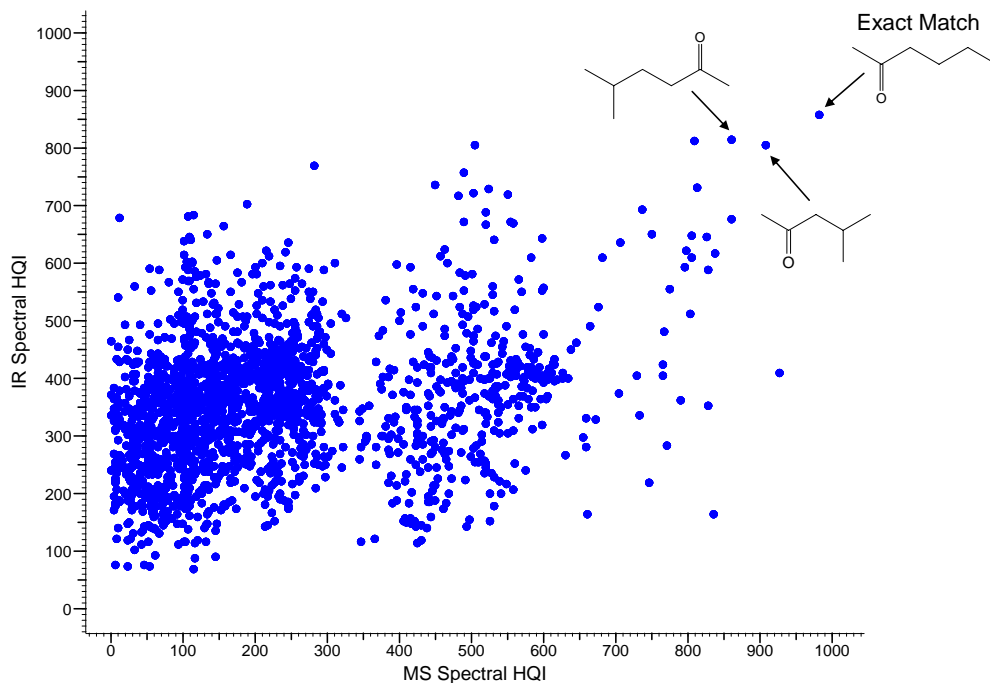


Figure 17 – Example 4 Results: 5,715 Spectra Pairs (MS/IR)

Conclusions

From this study, the following conclusions can be drawn from the fast and easy-to-use system of multi-technique spectral searching:

- Multi-dimensional analysis of multi-technique spectroscopic data is a simple yet incredibly powerful analytical technique for visualizing spectral space.
- The technique allows for efficient visualization of more data than is possible using a traditional spreadsheet approach, making this an effective datamining tool.
- The applicability of the technique in compound verification, unknown identification, and structure elucidation is clear.

For Additional Information

For additional information on the multi-technique spectral searching option of Bio-Rad's KnowItAll Informatics System, please visit our web site at:

www.knowitall.com